2 Introduction to Transducers and Sensors

2.1 Definitions of Transducer, Sensor, Actuator and Detector

In recent decades, advances in physics and electronics have enabled the development of devices that take information from a physical or chemical phenomenon and create or modify an electrical signal upon which this information is "copied". These devices are known by different names in instrumentation literature making unavoidable an introduction to some terminology frequently used.

In the electronic instrumentation literature some authors use the words transducer, sensor, actuator and detector in a way that could confuse the reader. This practice is often correct, because a device could meet more than one definition, but this is not always the case. Formal definitions from dictionaries provide some guidelines about how to use these words.

The Webster's on-line dictionary (http://www.merriam-webster.com/dictionary) defines **transducer** as:

> *"A device that is actuated by power from one system and supplies power usually in another form to a second system (a loudspeaker is a transducer that transforms electrical signals into sound energy)"*

The same on-line dictionary defines **sensor** as:

> *"A device that responds to a physical stimulus (as heat, light, sound, pressure, magnetism, or a particular motion) and transmits a resulting impulse (as for measurement or operating a control)"*

Dictionary definitions associate **transducer** with a device that converts one form of energy (or power) into another, and **sensor** with a device that perceives a physical stimulus giving a signal as a result. For example, microphones and hydrophones convert vibration into electricity, and thermocouples transform temperature into electricity, therefore they are examples of transducers (Fig. 2.1a).

Pressure sensors, resistance temperature sensors, thermistors, strain gauges and photoresistors are examples of sensors, because they are supplied with electrical energy and give an electrical signal when subjected to a stimulus (Fig. 2.1b).

Other authors (http://digital.ni.com) prefer to call **active transducers** those that generate an electric current or voltage in response to environmental stimulation, and **passive transducers** those that produce a change in some passive electrical quantity, such as capacitance, resistance, or inductance as a result of stimulation.

Generally, the word **actuator** refers to a device that converts an electrical signal into a mechanical motion. For example, an electric motor fits this definition because when powered by voltage produces a mechanical rotation of an axis. Also, the automatic locking of the car's doors is made by an actuator.

In the case of a car's horn, the electric circuit produces a mechanical vibration (motion) which is transferred to the environment (the sound), and so the horn could be said to be an **actuator**. Also, because the horn converts electrical energy into sound (mechanical energy) it might also be called a **transducer**. But it is not a **sensor** because it does not produce an electrical signal as a response to a physical stimulus.

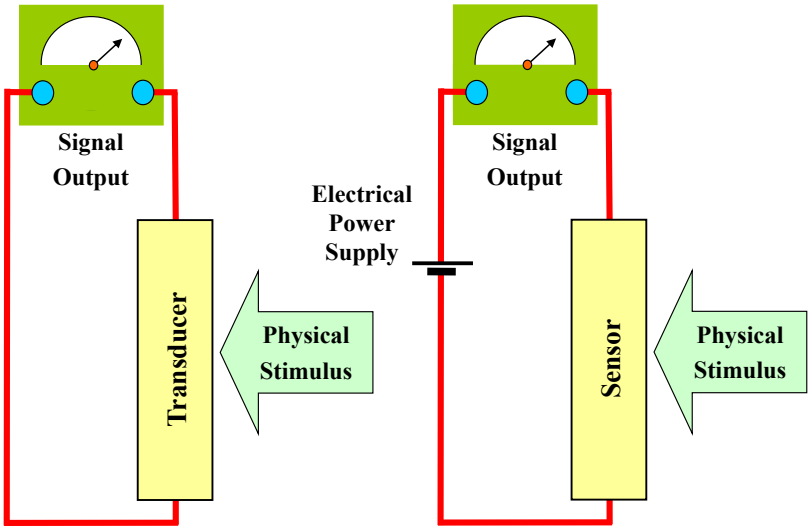

Fig. 2.1: (a) A transducer generates electrical energy from a physical stimulus. (b) A sensor "modifies" the electrical energy provided by the power supply.

To add more confusion, some articles use the term **detector**, defined as a device that recovers information of interest contained in a physical phenomenon and "impresses" it on an electrical signal. Therefore, it is similar to, and sometimes indistinguishable from, the concept of **sensor**.

These definitions are, however, disregarded or used interchangeably in some instrumentation literature, so the reader should be warned about these misuses of the words **sensor** and **transducer** just to avoid being confused. For example, photoelectric cells modify their electrical output signal when illuminated, so they are referred to as optical **detectors** or optical **sensors**. By means of the photovoltaic effect these devices transform the photon energy into electrical energy, and then they are also called **transducers**. As it happens with this example, there are many devices which are referred to in several ways, but it is useless to discuss each case. It is enough that the reader knows that one device can be referred to in different ways. In order to avoid any semantic discussions, the words sensor and transducer will be

preferentially used in this work. Therefore, do not worry about names. More important than adopting one or another name for a concept, is to understand it. Henceforth, our efforts will be focused in explaining the concepts.

2.2 Transfer Functions

The intrinsic function of a sensor is to produce an output in response to a stimulus. Therefore, in order that the sensor could have a practical use, that is to say, that the output could be used to measure the stimulus which produced it, the relation between input and output should be time invariant and well known. The stimulus (or the parameter which is being measured) will also be referred to as sensor input or measurand.

The transfer function (or transference) is a curve representing the relationship between the input and the output. In general, the input is the physical quantity being measured (physical variable, **Inp**) and the output is the magnitude of the electrical output (voltage or current, **Out**). Examples of these relationships are shown in Figure 2.2. The transfer functions of sensors are thus mathematical representations of the relation between their inputs and outputs.

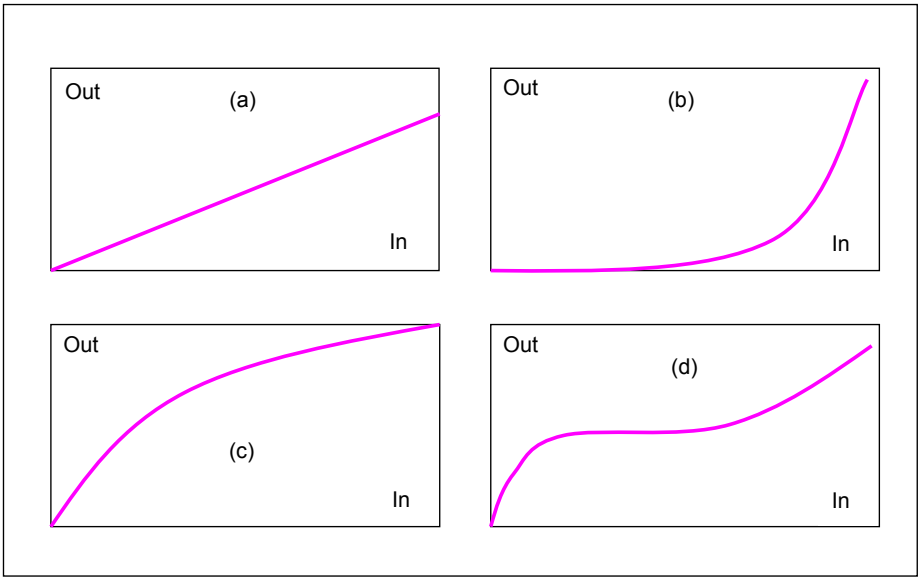

Fig. 2.2: Transfer functions: (a) linear transfer; (b) low sensitivity for small inputs; (c) low sensitivity for large inputs, (d) low sensitivity for middle inputs.

A transfer function may be regarded from a static or dynamic point of view. The static transference is the relation between the input and the output when the input is constant (does not vary with time). The dynamic transference describes the same relationship when the input varies with time. In other words, the output of a sensor can vary in a predictable way for different frequencies of the input; this idea is clarified below. We will first analyze the characteristics of static transfer functions.

Some real systems have linear input/output characteristics, so the representation is simply a straight line (Fig. 2.2a). The slope of a static transfer function curve at any point is called the **sensitivity** at that point. In measuring instruments it is generally convenient to have linear transfer functions because the sensitivity is constant, and the magnitude of the physical phenomenon being measured is obtained by multiplying the electric output by the reciprocal of the sensitivity, which is also constant.

The transfer function of Figure 2.2b shows a great sensitivity for large input values, but it is almost insensitive to small input values. Figure 2.2c, in contrast, shows a sensitive transfer function for small input values, but the sensitivity decreases as input increases. The curve of Figure 2.2d is insensitive for middle input values.

In Section (1.6) it was explained that it is desired to have great signal to noise ratios (*S/N*). Thus, for a given noise it is good to have the maximum possible signal. In low sensitive areas of the transference, sensors produce small electrical signals. So in these regions, electrical noise is more likely to interfere with measurements. When the sensor works in the high sensitive part of the curve it produces better data quality because for the same change in the measurand the output change is greater than in the low sensitive part. Therefore the *S/N* ratio increases, and the undesirable noise perturbs less the desired signal. Therefore, from the *S/N* ratio point of view, in a nonlinear transference the low sensitive part should be avoided.

The same concept of a transfer function is also applied to a complete instrument. In this case it is the relation between its output and input, the input being the measurand, and the output, the data measured, recorded or transmitted. As it was mentioned in the description of a generic instrument, an instrument has several stages connected in series, so the complete transfer function of the instrument can be considered as a set of partial transfer functions, also connected in series (Fig. 2.3). In this figure two main transfer functions can be identified, one corresponding to the sensor and the other to the electronics. Figure 2.3 shows how the transfer function of the electronics may be tailored to compensate the sensor's nonlinearities, giving as a result a linear transference for the complete instrument. The electronics include both analog and digital parts; the linearization process may be performed in either part or both.

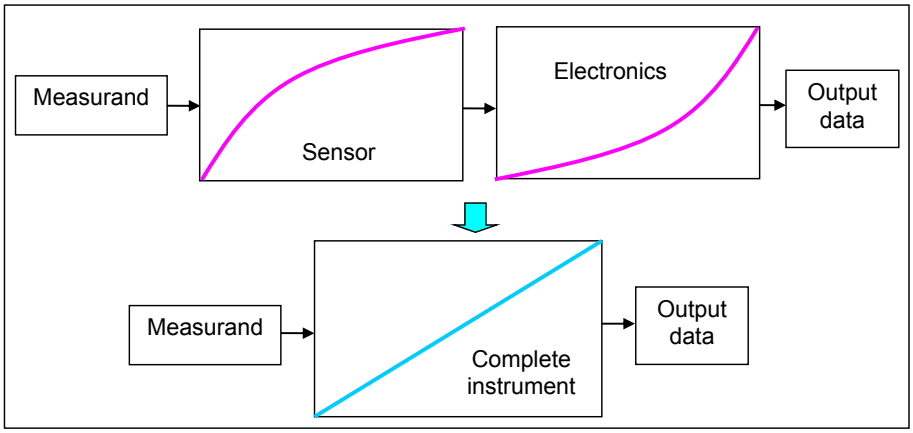

Fig. 2.3: In the upper part, the electronics' transference is designed to compensate the non-linear sensor's transference, giving as a result a complete linear transference (lower part).

2.2.1 Range and Dynamic Range

The input **range** is the difference between the maximum (X_{max}) and minimum (X_{min}) values of the measurand that can be measured by the instrument. The dynamic range (D_{range}) is the ratio between the largest and the smallest measurable signal.

$$Range = X_{max} - X_{min}; \quad D_{range} = \frac{X_{max}}{X_{min}} \tag{2.1}$$

For example, assume that the conductivity of pure water is 0.5 μS/cm and that of seawater is about 50 mS/cm, then if we want to have one conductivity meter to measure from pure water to seawater it should have a dynamic range of 100,000 (also found in manufacturer's brochures as 1:100,000). It is very difficult to cover such a range keeping the measuring error low. So with the purpose of having low errors, conductivity meters are manufactured to measure in a much limited dynamic range, for example 1:1000. Other kinds of instruments have lower dynamic ranges. For example, some electromagnetic current meters can be used only in a dynamic range 1:50, if errors below 0.5 % are desired; vortex flowmeters are insensitive to low velocities, and can be used in the dynamic range 1:10.

Range and dynamic range have been defined for the measurand (input signal) but these definitions also apply for the output of instruments.

2.2.2 Hysteresis

A simple example will be useful to introduce the concept of hysteresis. Some synthetic elastic materials stretched by a force do not return to their initial length when the force is released. Then if a two-dimensional graph is constructed using several pairs of length and force points, two curves will be observed, one when the force is increasing and another when the force is decreasing. It is said that the rubber does not follow Hooke's law because it has some hysteresis.

Likewise, there are sensors whose transfer functions depend not only on the existing input signal but on the previous states of the input (Carr & Brown, 1998). This means that there is not a unique curve (transfer function) but it depends on input signals increasing or decreasing. Thus, the transfer function tends to form a loop as in Figure 2.4. The **hysteresis** (H) is evaluated by means of the maximum variation (h_{max}) observed in the output of a transfer function produced by upscale (increasing) and downscale (decreasing) input variations. It is defined as a percentage of the input range (Eq. 2.2).

$$H\,(\%) \;=\; \frac{h_{max}}{Range} \times 100 \tag{2.2}$$

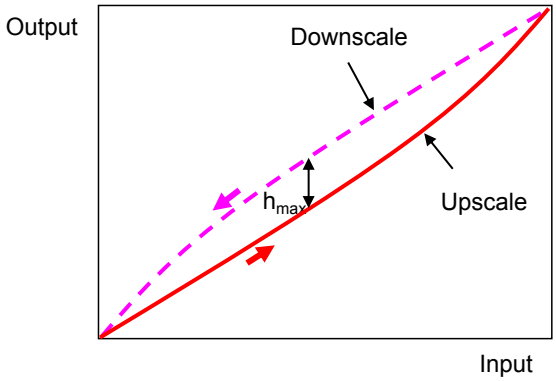

Fig. 2.4: Hysteresis. Upscale and downscale transferences; h_{max} is the maximum variation between both.

2.2.3 Calibration

The word **calibration** may be used in slightly different ways depending on the context. In the vocabulary used by quality control professionals, calibration is the periodically repeated process for verifying the capability and performance of a measuring instrument by comparison to traceable measurement standards. The calibration process assures that the instrument remains capable of making

measurements according to its performance specifications. It consists in comparing a given instrument with a measurement system or device having higher specifications and known uncertainty. This comparison allows any variation in the accuracy of the instrument under test to be detected.

It is easily understood that a new instrument could need an initial calibration just after being manufactured, so the word calibration is also used to call the process of alignment or adjustment of an instrument when it is either manufactured or repaired. In order to determine the transfer function of a just manufactured or repaired instrument, other instrument of better quality or a standard instrument has to be used.

For the case in which the sensitivity is variable (nonlinear transference), it may be necessary to determine the transfer curve point by point, which is often a slow and laborious work. This is another reason why sensors with linear transfer functions are more convenient; it is (theoretically) possible to obtain their transfer functions with only two calibration points.

2.2.4 Linearity

The linearity of a sensor (or an instrument) expresses how much its actual transfer curve deviates from a straight line. It is evaluated by the percentage of non linearity as defined by

$$NL\ (\%) = \frac{D_{\max}}{Range} \times 100 \tag{2.3}$$

where NL (%) is the percentage of nonlinearity, $D_{\max}$ is the maximum output deviation from the straight line and **Range** is the total output range = $O_{\max} - O_{\min}$.

The percentage of non linearity depends on how the linear approximation of the curve is done (Carr & Brown, 1998; http://www.sensorland.com). From Figure 2.5 the maximum output deviations *D1* and *D2* are different, thus nonlinearity substantially depends on how the straight line is drawn. When the **linearity** of an instrument is specified by its manufacturer, a statement on how the straight line is superimposed on the transfer curve should be made. The user of instruments should be aware that manufacturers specify the possible minimum nonlinearity to show the goodness of their instruments and, in some cases they forget to explain what method they used to draw the straight line. Users should ask manufactures how they specify linearity. Otherwise the information becomes meaningless.

One usual way to make a linear approximation of the transference curve is by means of a straight line whose slope is calculated from the two end (or terminal) points of the transfer function, as shown in Figure 2.5. It is known as the Terminal Straight Line (TSL) method. In this method, the manufacturer calibrates a zero point and a full scale point. The maximum output deviation is *D1*; nonlinearity errors are the smallest around the end points and the highest somewhere in between (Fig. 2.5).

Another way to calculate the linear approximation is by using the Least Squares Best Fit Straight Line (BFSL) method. It is a statistical method that requires knowledge of several calibration points over the whole working range of the sensor. The input and output values at each point are then used to calculate the slope of the straight line which provides the closest possible best fit to all data points on the curve and the intercept. Figure 2.5 shows that the maximum output deviation *D2* is smaller than in the TSL method.

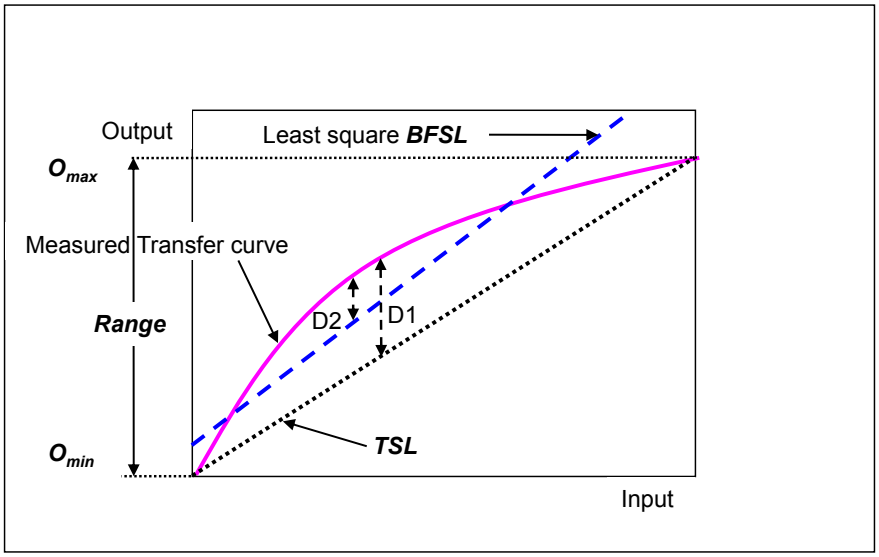

Fig. 2.5: Non-linear transfer curve with TSL and BFSL linear approximations. $D1 \neq D2$ shows that the definition of linearity depends on the linear approximation method.

2.2.5 Offset and Gain Errors

Due to various causes such as daily usage, shocks, vibration, etc., an instrument could suffer changes in its transference, and some measurement error may thus arise when using it. In an instrument with a linear transference, two kinds of changes from the initial calibration can be found; these departures can produce some measuring errors which are known as **offset** and **gain errors**.

The **offset error** of a sensor (or an instrument) is defined as the output that exists when the input is zero. It could be thought as a shift of the whole transfer function by a constant amount. In order to measure the offset a zero input must be forced; for example, in a flowmeter the instrument must be removed from the flow or the

flow must be blocked. Then, the output must be read and if the offset were invariant with time, the offset error could be removed by subtracting it from the actual transfer function (Fig. 2.6).

If the slope of the actual transference differs from the ideal (or calibrated) transference it is said that the instrument has a **gain error** (Fig. 2.6).

So far we have allotted the departure from the ideal transfer curve of equipments to their use, but a new instrument also has some offset and gain errors because there is a limit in the quality of the manufacturing processes of sensors (and instruments), which generates a transfer function somewhat different from the ideally desired. In general, manufacturers specify the ideal transfer function as if all sensors were made equal, but at the same time, they also specify the offset and gain errors as a percentage of the total output range or as a percentage of the reading.

Furthermore, if the instrument were subjected to aging, manufacturers should specify the percentage change of **offset** and **gain** per unit time. This happens with some sensors whose transfer characteristics are based on a sensitive material whose properties vary with time in a well known way, and so the change in the transference can be predicted. This change is expressed, for example, as a given gain change of -1% per year for a sensor that is loosing its sensitivity.

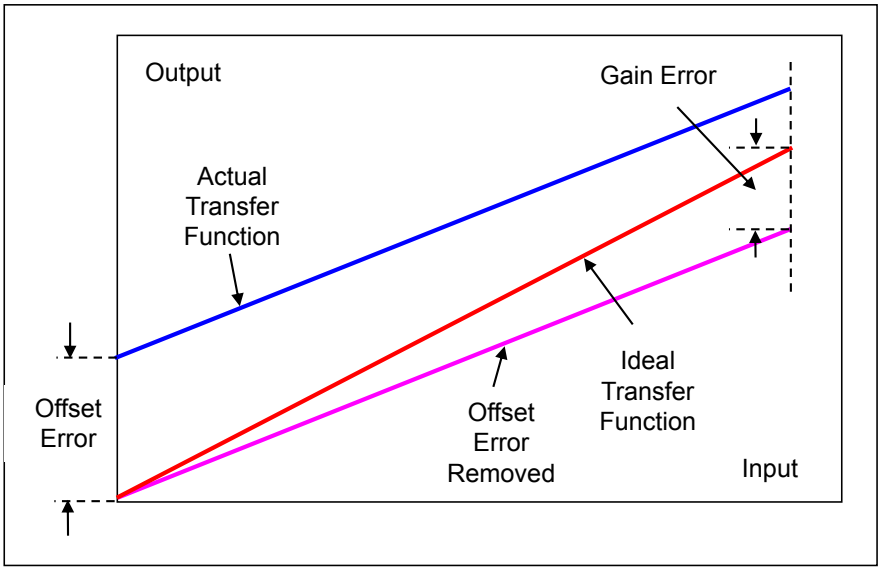

Fig. 2.6: Actual transfer function with offset and gain errors; ideal transfer function; and transfer function with offset error removed. At null and full input, offset and gain errors are indicated.

2.2.6 Drift

Real sensors are not time-invariant systems. Once offset and gain errors are eliminated by some kind of electronic compensation, the transfer function could drift due to environmental factors such as temperature and humidity. The manufacturers usually specify the **zero drift** and the **span drift** as a function of the factor that modifies the transference. For example, for an instrument that measures water table from zero to 10 m a **zero drift** = - 0.1% /°C of full scale, means that for an increase of 10 °C in the instrument's temperature, the measured value will decrease 10 cm, which is not negligible for some environmental studies.

2.2.7 An Example of Sensor Specifications

As an example, the following parts of the whole specifications of a commercial pressure sensor were chosen:

Accuracy ..... 1.0% Full Scale
Non-Linearity and Hysteresis ..... 1.0% Full Scale
Non-Repeatability ..... ± 0.15% Full Scale

Accuracies stated are expected for Best Fit Straight Line for all errors, including linearity, hysteresis & non-repeatability **(Note 1).**

Temperature, Operating ..... 50° to 145°C
Temperature, Compensated ..... 15° to 70°C
Temperature Effects
Zero drift ..... 0.02% Full scale/ °C
Span drift ..... 0.04% Reading/ °C **(Note 2).**

Special Calibration
10 point (5 up/ 5 down) 20% increments @ 20°C
20 point (10 up/ 10 down) 10% increments @ 20°C **(Note 3).**

Comments to the specifications

Note 1: This manufacturer does not specify each error, but gives figures that include all of them. Also, the manufacturer tells us which was the method used to linearize the transfer function (Best Fit Straight Line).

Note 2: The sensor can be used between -50° and 145 °C, but specifications are met only for the compensation range between 15° and 70°C. Temperature changes affect

the transfer function. The zero drift is specified as a percentage of the full range, and the span drift as a percentage of the measured value (reading).

Suppose that the full range of this sensor is 10 m of water column (wc) and that it will be used to measure the height of a water column in a tank in an environment whose temperature is limited to the interval between 15 and 70 °C. Let us assume that we calibrate the sensor at 15 °C (offset and gain adjusted). If when performing the measurements the temperature of the water is 70°C, the zero drift for the total change in temperature (70°C – 15°C = 55°C) will be

$$\textit{zero drift} = 0.02\%\ \textit{Full Scale} / {}^{\circ}\mathrm{C} = \frac{0.02 \times 10\ \mathrm{m} / {}^{\circ}\mathrm{C}}{100} \times 55{}^{\circ}\mathrm{C} = 0.11\ \mathrm{m}$$

If the instrument is measuring at a height of 5 m (half of the total range) the span drift will be:

$$\textit{span drift} = 0.04\%\ \textit{Reading} / {}^{\circ}\mathrm{C} = \frac{0.04 \times 5\ \mathrm{m} / {}^{\circ}\mathrm{C}}{100} \times 55{}^{\circ}\mathrm{C} = 0.11\ \mathrm{m}$$

Then, the total drift would be 0.22 m over a range of 5 m, which is a total change due to drift of 4.4 % for a change in sensor temperature of 55 °C. Obviously it seems very difficult to find environmental measurements where the fluid changes its temperature in such large amount; but it could be quite reasonable in an industrial process.

This somewhat unrealistic example was chosen to show the importance of temperature calibration. If the temperature of the water whose height has to be measured would range from 15 to 70 °C, it would be much better to calibrate the pressure sensor at half the range = 42.5 °C. In so doing, the change due to drift will be ± 2.2 %.

It should be noted that if the tank had a small column of water of, say, 1 m the zero drift would be still 0.11 m and the span drift would reduce to 0.022, the total drift being 0.132 m. But the total change due to drift is 13.2 % of the measured value. Therefore, an important conclusion is that to minimize the error due to zero drift, the sensor has to be selected in such a way so that it does not measure in the lower part of its range.

This is a good example to show that a sensor has to be selected according to the desired measuring range with the purpose of minimizing errors and, also, to keep a good signal to noise relation. Whenever possible, sensors should be selected to measure in the upper part of their ranges. It is obvious that a sensor able to measure 100 m water column will measure 1 m too, but the price paid is that noise and errors can make the measurements useless.

Note 3: This manufacturer offers two kinds of calibration (at different costs): the 10 points and the 20 points. The first means that the Best Fit Straight Line is obtained from 5 points increasing input values and 5 points decreasing input values. Obviously, the 20 points calibration (10 increasing and 10 decreasing) helps to better know sensor errors due to non-linearity, but the price of the sensor will be higher.

2.3 Spatial Characteristics of Sensors

2.3.1 The Decibel

In order to understand some properties of sensors it is necessary to introduce a useful way of comparing two quantities; e.g. in some cases it is convenient to compare them on a logarithmic scale. This idea was first extensively used for telephone power line (P) calculations. The unit of this logarithmic scale is called the decibel (dB). Given two power values P_2 and P_1 the number (N) of decibels is defined by

$$N = 10 \log \frac{P_2}{P_1} \tag{2.4}$$

A negative value of N means that $P_2 < P_1$.

This way of comparison between two power values was then extended to other fields such as sensor theory, in which it is used to compare two signal amplitudes (A). The relative change in signal amplitude of sensors is commonly measured in decibels. Because P is proportional to the square of the voltage amplitude ($A2$), when comparing two amplitudes the number (N) of decibels by which $A2$ exceeds $A1$ is defined by

$$N = 10 \log \left(\frac{A_2}{A_1}\right)^2 = 20 \log \frac{A_2}{A_1} \tag{2.5}$$

Since N compares two quantities, $A1$ and $A2$, in a relative way, in order to get the absolute value of one of them it is necessary to know the other. For example, when it is said that the voltage gain of an amplifier is 80 dB it means that two voltage **amplitudes** are compared and Eq. (2.5) must be used. From this equation it is found that 80 dB corresponds to an amplification of 10,000 times, but the absolute value of the output voltage remains unknown. If the reference were one millivolt (1 mV) the voltage output would be 10 V.

Obviously, the decibel may be used to give absolute values when a reference value is adopted. Sometimes a suffix is added to the decibel symbol to indicate which reference has been used for comparison. That is the case of dBm, which indicates that one milliwatt (1 mW) has been use as the reference value. A sound amplifier having a 40 dBm gain will have an output of 10 W. Note that because dBm indicates that the reference is a unit of *power*, Eq. (2.4) must be used.

Briefly, when the instrument specifications are expressed in dB, to know absolute values it is required first to recognize if quantities compared are *amplitudes* or *powers*, and second, what is the *reference* value.

Some useful approximated relationships are displayed in Table 2.1.

Table 2.1: The decibel. Approximated relationships for power and amplitude

Ratio	Log_{10}	Power (dB)	Amplitude (dB)
1	0	0	0
1.414	0.1505	1.5	3
2	0.301	3	6
3.16	0.4997	5	10
5	0.6989	7	14
10	1	10	20
100	2	20	40
1000	3	30	60
10000	4	40	80
100000	5	50	100
1000000	6	60	120

2.3.2 Sensor Directivity

A very important property of a sensor is its directivity. Sensors were previously associated with devices that perceive a physical stimulus giving as a result a signal (as explained above, in most devices the output is an electric signal). So far nothing was said about the relative position between the sensor and the direction of the stimulus. A device is said to be **omnidirectional** if it produces the same output signal regardless of the direction where the stimulus comes from. Figure 2.7 shows the voltage output signal of an omnidirectional sensor as a function of the arriving direction in polar coordinates (scale must be multiplied by 10 to cover 360 °). The radial axis is the output voltage (7 V in this case). Few practical sensors are omnidirectional; most of them produce a higher output signal when the stimulus arrives from some preferential direction. This property is often used to know where the signal comes from.

One way to specify the spatial sensitivity of a sensor is to plot its directivity pattern. Usually, the three-dimensional pattern is plotted in two planes containing the axes of the device (generally the horizontal and vertical planes); plots are typically in polar coordinates (Figs. 2.8 and 2.9). The polar diagram of Figure 2.8 shows the main lobe (upwards) and the two secondary lobes (downwards). The main lobe gives the direction of the maximum sensitivity of the sensor, or the maximum radiation angle in the case of an emitting transducer (e.g., a radar antenna). Commonly the main lobe is pointed towards the stimulus of interest. It is desired that the difference expressed

in dB between the main and the secondary lobes be as large as possible, because the secondary lobes could catch signals from undesirable stimuli, thus masking or compromising the desired measurement.

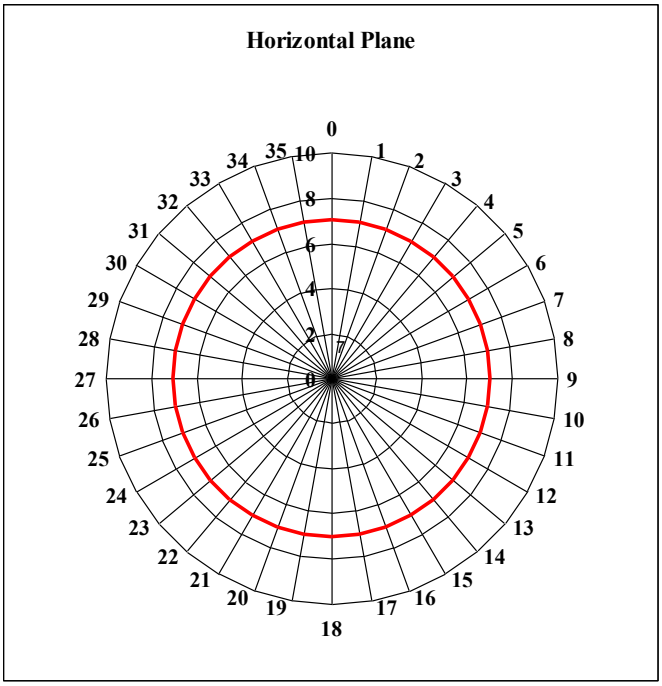

Fig. 2.7: The output voltage of a sensor as a function of the direction the stimulus comes from represented in the horizontal plane. The scale shown on the perimeter must be multiplied by 10.

In the examples shown in Figures 2.8 and 2.9 the horizontal pattern is more directive than the vertical one. A measure of the directivity of sensors is given by its **beamwidth**. It is a measure of the angle between two points of the lobe at some level below the maximum sensitivity angle. Some of the most commonly used levels are -3 dB, -6 dB and -10 dB but manufacturers can define their own.

Generally, two powers are compared in defining the **beamwidth** and so Eq. (2.4) must be used. The level used to define the beamwidth must be specified together with the beamwidth. For example, taking the horizontal pattern of Figure 2.8, the maximum level is $P1 = 10$ and choosing the reference level of -3 dB ($P2/P1 \approx 0.5$), on the left side of the main lobe this level corresponds to about 335 ° (or -25 °), and on the right side it is about 25 °. Then the beamwidth for -3 dB is 50°.

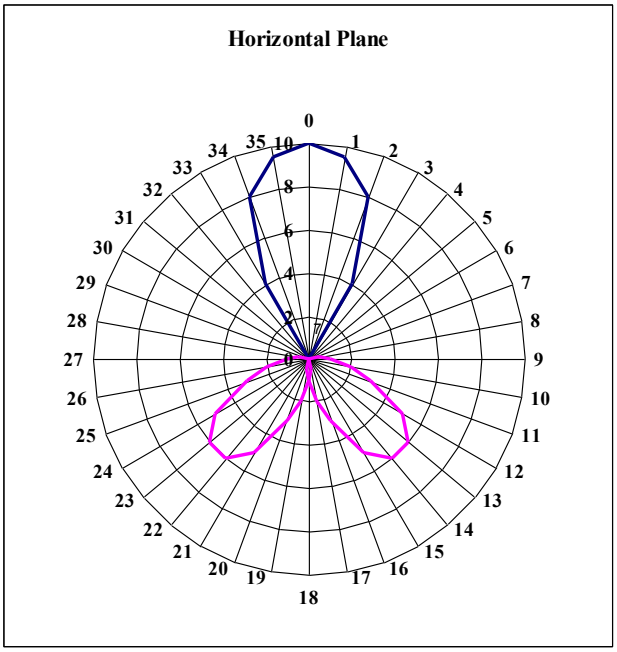

Fig. 2.8: Horizontal plane beamwidth for -3 dB ($P_2/P_1 \approx 0.5$), is - 25 ° + 25 °.

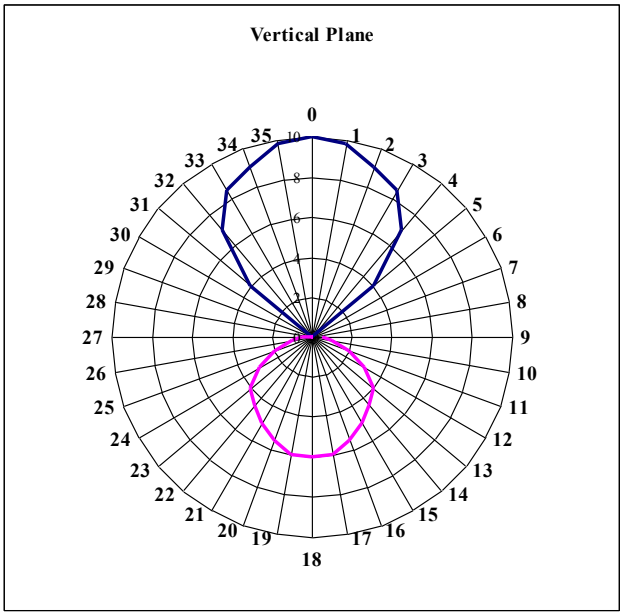

Fig. 2.9: Vertical plane directivity pattern.

2.3.3 Spatial Averaging

The concept of spatial averaging is introduced here to make readers note that measured values do not correspond to an infinitesimal point on the space but to some volume. This volume should be known and compared with the spatial signal we want to evaluate. Suppose that an acoustic sensor for measuring waves is placed on a coastal structure at a height R above sea level with its main lobe pointing down. Assuming that the measurement zone is the intersection of the beamwidth with the sea surface, the area of the sea from which the sensor would take the information is the base of the cone illustrated in Figure 2.10,

$$S = \pi r^2 = \pi \left(R \tan\left(\frac{\alpha}{2}\right)\right)^2 \tag{2.6}$$

where r is the base radius of the cone and α is the vertex angle. For a sensor with a conical beamwidth of $\alpha = 30°$ (for a -3 dB level in both the horizontal and vertical planes) and placed at $R = 10$ m, $r = 2.68$ m and the surface of the cone base is $S = 22.6$ m^2. This means that the electrical signal generated at the sensor output will contain information from the stimulus contained in that area. The sensor is thus performing a spatial averaging of the wave information. Waves whose wavelengths (λ) are shorter than the diameter of the cone's base ($\lambda < 2r = 5.36$ m), will be averaged giving a reduced output.

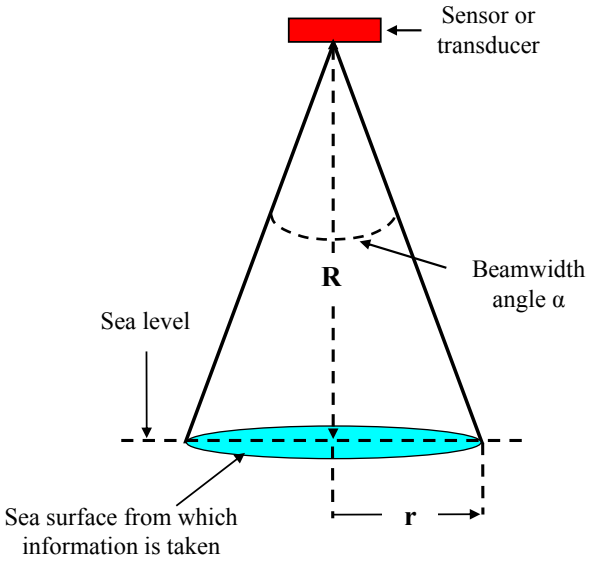

Fig. 2.10: The beamwidth of and acoustic wave sensor takes information of the sea surface averaged over the base of the cone.

If higher spatial resolution is needed (e.g., if it is desired to measure waves of short wavelength), the height of the sensor must be decreased, or a sensor with a narrower beamwidth adopted. For example, to measure the information contained in an area $S < 1\ m^2$ ($\lambda < 2\ r \approx 1.13$ m) the sensor should be within 2.1 m height above sea level. By the way, in some locations where the tidal range is significant, the wave spatial averaging of such an acoustic wave meter will be a function of the tidal level, which could be inconvenient for some applications.

This phenomenon of spatial averaging occurs for most sensors. Therefore, to recognize whether the sensor used is actually measuring the desired phenomenon, it is important to assess its spatial averaging.

2.4 Time and Frequency Characteristics of Sensors and Systems

2.4.1 Introduction

So far we have analyzed the **static transfer function** of sensors and instruments. We will now turn our attention to the **dynamic transfer function**, which describes how sensors' inputs and outputs are linked when the input signals vary in time. Although this topic is included in this chapter devoted to sensors and transducers, the concepts to be developed are very general and applicable to instruments and systems.

All devices have a finite capacity to reproduce at the output any time-varying signal applied to the input. This means that the time variations of measurands are modified somewhat by sensors and instruments, and users should evaluate if these modifications are significant for the phenomena that they are studying.

Manufacturers specify the capacity of the sensor or instrument to reproduce the measurand dynamic behaviors in a variety of ways. They define different parameters that account for these characteristics: frequency response, bandwidth, time constant, rise time, etc. It is important for the users to understand the meaning of these specifications to evaluate whether any particular instrument satisfies their measuring requirements. As can be intuited from their names some specifications refer to the frequency domain and some to the time domain.

A clear understanding of the above concepts will be useful to users in order to check whether they are losing part of the desired measurand information due to dynamic limitations of sensors or instruments. Also, in those cases that manufacturers do not provide such parameters to account for the instrument dynamic response, these concepts would enable users to perform their own experiments to estimate them. These tests could be as simply as applying a step excitation to the instrument input. This kind of check could become healthy routines before field installation of instruments when some doubts exist about the dynamic instrument performance or one of the following situations occur: factory specifications are not available; the instrument has been used by a prolonged time without maintenance; the instrument has been a prolonged time in storage without use; or the instrument has undergone a deep repair.

2.4.2 Frequency Content of Signals

The **period** (T) of a signal is the elapsed time between two repeating events. The f**requency** (f) is the number of occurrences of the events per unit time, e.g. the frequency of the heart beat is about one beat per second = 1 Hz. Thus, frequency is the reciprocal of period; $f = 1/T$. The SI unit for period is the second (s) and for frequency is the hertz (Hz), the Hz being 1/s.

Most waveforms of practical use in engineering can be analyzed by studying their energy content at each frequency. If the waveform is periodic (i.e. if there is a positive nonzero value of T such that the value of the function at time $t + T$ is equal to its value at time t), it can be described by a series of sine and cosine terms whose frequencies are integer multiples of a fundamental frequency (f_0), which is the inverse of the (fundamental) period (T_0) for the waveform to repeat itself. This series is called a **Fourier series**. The Fourier theorem states that any periodic signal, no matter how complex it is, may be seen as a combination of a number of pure sine and cosine terms with harmonically related frequencies.

If the wave is not periodic, the fundamental period tends to infinite ($T_0 \rightarrow \infty$). The fundamental frequency ($f_0 = 1/T_0$) is therefore infinitesimal and the frequencies of successive terms of the Fourier series differ in an infinitesimal amount instead of in a finite amount. The Fourier series becomes thus the **Fourier integral**. It turns out that any waveform can be decomposed or analyzed into only sine and cosine functions of different amplitudes and frequencies.

A fundamental concept already intuitively introduced should be clear to researchers: **all sensors, and then instruments, deliver a signal that is a modified version of the real signal they want to measure**. The measuring process changes the temporal (and frequency) characteristics of the signal, sometimes in a seamlessly, and therefore acceptable way, and others in an appreciable and inconvenient way. This undesirable trimming of the signal is due to an unwanted filtering which restricts the frequency content of the input signal. This happens because real devices cannot respond to very fast or very slow stimuli. Figure 2.11 shows how a signal could be distorted due to unwanted filtering. The amount of distortion and attenuation of the parameters being measured depends on the relation between **signal frequency content** and **instrument frequency response**.

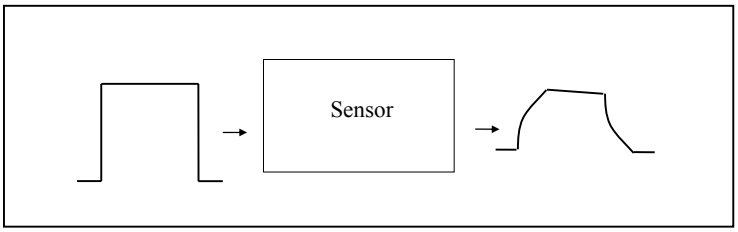

Fig. 2.11: The input signal (left) is modified by the sensor frequency response giving a different signal at the output (right).

2.4.3 Frequency Response

Frequency response is one of the ways in which manufactures inform the users how instruments perform dynamically. The **frequency response** of a device describes its ability to reproduce at the output any signal applied to the input. The **frequency response** of a system is a measure of the magnitude and phase of the output as a function of frequency, in comparison to the input.

In simple terms, when a sine wave of given frequency is injected at the input of a linear system, it will output that same frequency with a certain magnitude and a certain phase angle relative to the input. To obtain the **frequency response** of a system, sinusoidal inputs [$x(t)$] of known amplitudes may be applied to the system, and the corresponding output amplitudes and phases (relative to the inputs) [$y(t)$] must be measured over a range of frequencies,

$$x(t) = X \sin\ 2\pi f_0 t; \qquad y(t) = Y \sin(2\pi f_0 t + \phi) \tag{2.7}$$

The corresponding gain $G(f_0)$ and phase $\varphi(f_0)$ of the frequency response for the frequency f_0 are defined as

$$G(f_0) = \frac{Y}{X}; \qquad \varphi(f_0) = \phi \tag{2.8}$$

The following example shows how a real instrument can distort part of the information contained in the measurand. The approximate transfer function, gain and phase as function of f (*frequency response*) of an old commercial wave meter is shown in Figures 2.12a and 2.12b. These figures are also referred to as the **Bode plot** of the instrument response.

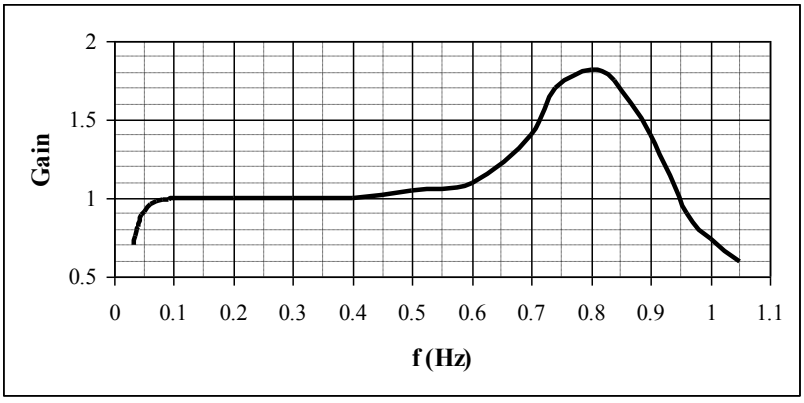

Fig. 2.12a: Wavemeter gain as a function of frequency. As modified from Datawell (1980).

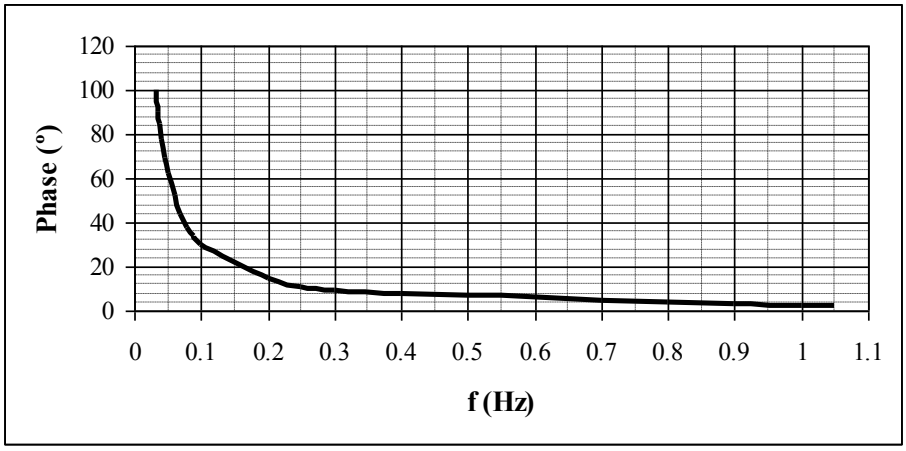

Fig. 2.12b: Wavemeter phase output relative to the input, as a function of frequency. As modified from Datawell (1980).

Let us see how the instrument modifies the recorded data. Waves of frequencies $0.1 < f < 0.4$ Hz (or period $2.5 < T < 10$ s) are measured without attenuation ($G = 1$) and with a small change of phase. Long waves of $f = 0.05$ Hz ($T = 20$ s) are attenuated by 10 %; short waves between $0.4 < f < 0.6$ Hz (or period $1.67 \text{ s} < T < 2.5$ s) are overestimated up to 10 %. Waves of $T = 1.25$ s are overestimated up to 80 %.

For most studies this instrument is adequate, but if the study requires to measure outside the range from 0.1 to 0.6 Hz perhaps another instrument best fitted to the desired range could be found. Then, before using an instrument the frequency range of the phenomenon to be measured should be compared to the transfer function of the instrument.

Manufacturers of the wavemeter whose Bode plots were presented did a good job specifying their instrument and allow users to better know how the measuring system influences the recorded data. Unfortunately, many times, manufacturers do not provide a complete Bode plot of their devices.

In some particular sensors and instruments whose transferences are more flat and regular than that shown in Figures 2.12a and 2.12b, manufacturers give the plot of the gain as a function of frequency, $G(f)$ and phase shifts are inferred from some equations accepted as standard. The transference region in which gain is constant permits the useful frequency range of the instrument to be identified, also known as the **bandwidth** of the instrument.

2.4.4 Bandwidth

Bandwidth is used to characterize electronic filters, sensors, antennas, audio amplifiers, instruments, communication channels, etc. It describes their ability to allow the passage of a signal without significant attenuation and distortion. Figure 2.13 shows the gain of a given device together with its bandwidth. The frequency band of a circuit or sensor in which the output signal is attenuated less than -3 dB is called the **bandwidth.**

In Figure 2.13 there is a range of frequencies over which the output $G(f)$ is almost constant (G_0) and then all the input information is translated to the output without amplitude distortion. At frequencies f_1 and f_2 the output falls to $G(f_1) = G(f_2) = 0.707\ G_0$; this drop in signal level corresponds to -3 dB. Frequencies f_1 and f_2 are called the lower and upper frequency points, respectively, and the bandwidth is defined as $B = f_2 - f_1$. For many practical cases $f_1 \approx 0$, so $B = f_2$.

The -3 dB points are also referred to as the half-power points because power is proportional to the square of signal amplitude, then a drop to 0.707 in amplitude means a drop in power to $P \equiv (0.707)^2 \approx 0.5$.

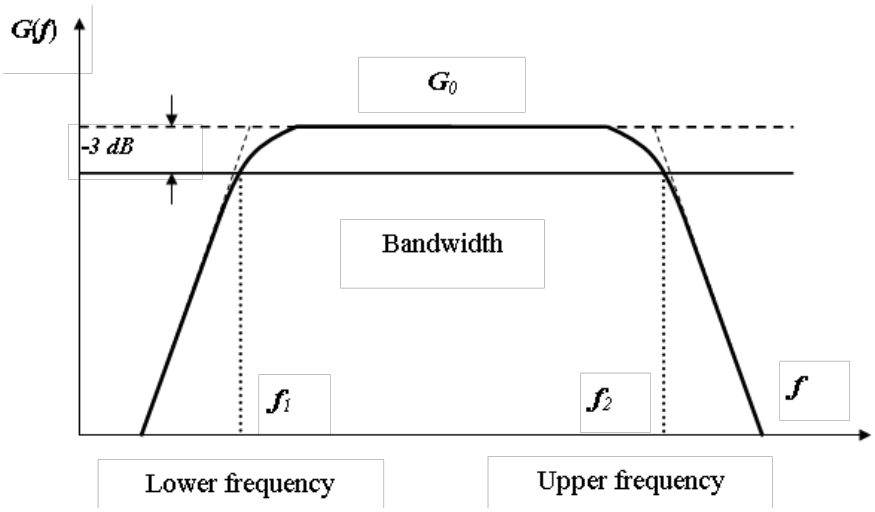

Fig. 2.13: Gain as a function of frequency showing a constant gain (G_0) over a certain range of frequencies. Lower and upper frequency points (f_1 and f_2, respectively) and the bandwidth are also shown.

There is a simple equation for the *bandwidth* of a first order linear time invariant (LTI) system. This simplification is very useful because in nature there are many first order LTI systems, among them thermal, hydraulic, mechanical, biological, radioactive

phenomenon may be treated as LTI systems. A similar behavior can be found in electronic circuits, viscoelastic materials, humidity meters, etc.

Out of the almost constant (G_0) region (central region of the transference of Figure 2.13 where the information is not modified) it can be demonstrated (Millman & Halkias, 1967) that for a sinusoidal forcing function, the corresponding gain ($G(f)$) and phase ($\Phi(f)$) of the frequency response of a first order LTI system are:

$$\begin{aligned} G_{UP}(f) &= \frac{1}{\sqrt{1+(N_2)^2}}; \qquad \Phi_{UP}(f) = \arctan\ N_2; \qquad N_2 = \frac{f}{f_2} \\ G_{DOWN}(f) &= \frac{1}{\sqrt{1+(N_1)^2}}; \qquad \Phi_{DOWN}(f) = -\arctan\ N_1; \quad N_1 = \frac{f_1}{f} \end{aligned} \tag{2.9}$$

where G_{UP} and Φ_{UP} are valid for high frequencies and G_{DOWN} and Φ_{DOWN} for low frequencies.

The gain and phase given by Eq. (2.9) are plotted as a function of $\boldsymbol{N = N_1 = N_2}$ in Figures 2.14a and 2.14b. For **N** = 1 the amplitude of the output is as expected from the definition of the upper and lower frequencies $G_{UP} = G_{DOWN} = 0.707$ times the input; and the phase is $\Phi_{UP} = +45°$ and $\Phi_{DOWN} = -45°$.

Typically it is desired that the input signal passes the system with minimum attenuation, and for this to happen, as a rule of thumb, the upper frequency of the system bandwidth should be one decade up the maximum frequency of the signal input and the lower frequency should be one decade down the minimum frequency of the input. For this condition, the relations of frequency should be $f/f_2 = f_1/f = 0.1$ resulting $G = 0.995$ (see Table 2.2 following) or, in other words, the attenuation would be only 0.5%.

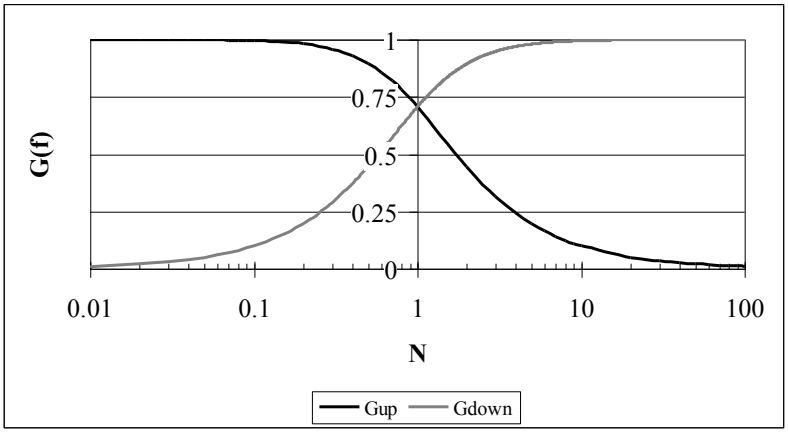

Fig. 2.14a: G_{UP} and G_{DOWN} are the gains in the upper and lower frequencies, respectively, as a function of $N = N_1 = N_2$.

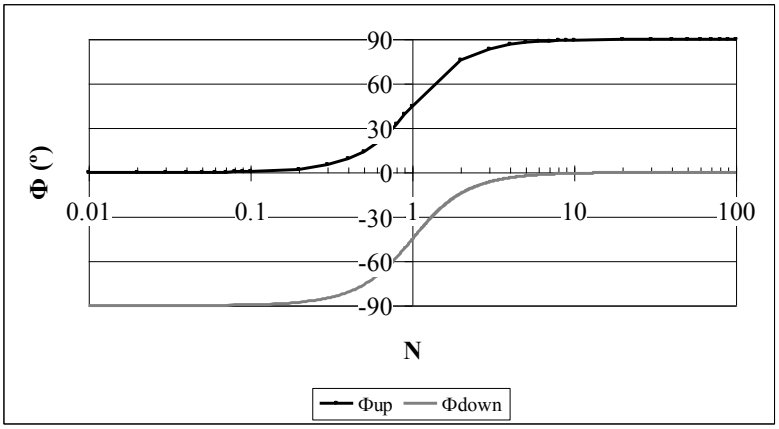

Fig. 2.14b: Φ_{UP} and Φ_{DOWN} are the phases in the upper and lower frequencies, respectively, as a function of $N = N_1 = N_2$.

Another less demanding rule of thumb states that the output signals attenuated and distorted by the system are still reasonably good if the frequency components of the input signal are included in the *bandwidth* of the device. That is to say that it is accepted that the measurand information content, close to the upper and lower frequencies, be somehow attenuated, because the output at f_2 and f_1 is 0.707 of the input. This means that a researcher who uses in the entirely bandwidth an instrument whose frequency response correspond to that of a first order LTI system will have the lower and higher frequency components attenuated about 30%.

Some points of Gain and Phase are shown in Table 2.2.

Table 2.2: Gain and Phase

N2	Gain	Phase (°)	N1	Gain	Phase (°)
0.1	0.99503719	0.5729387	10	0.99503719	- 0.5729387
0.2	0.98058068	2.29061004	9	0.98058068	- 2.29061004
0.3	0.95782629	5.14276456	8	0.95782629	- 5.14276456
0.4	0.92847669	9.09027692	7	0.92847669	- 9.09027692
0.5	0.89442719	14.0362435	6	0.89442719	- 14.0362435
0.6	0.85749293	19.7988764	5	0.85749293	- 19.7988764
0.7	0.81923192	26.104854	4	0.81923192	- 26.104854
0.8	0.78086881	32.6192431	3	0.78086881	- 32.6192431
0.9	0.74329415	39.0074726	2	0.74329415	- 39.0074726
1	0.70710678	45	1	0.70710678	- 45

Higher orders of **LTI** systems are also described by simple equations similar to Eq. (2.9) but it is left to the interested readers to find them in references such as (Millman & Halkias, 1967).

2.4.5 Time Constant

Hitherto the discussion has been centered in how manufacturers specify the capacity of instruments to reproduce the measurand dynamic behavior based on characteristics defined in the frequency domain. Because instruments record data as a function of time it would be interesting to find some parameter in the time domain that could give us an idea of how instruments modify the dynamic characteristics of the measurand. It will be even better if this time measured parameter could be related to the bandwidth of the system. This is difficult to do with complex transferences such as that of the wavemeter of Figures 2.12a and 2.12b, but simple approximations could be proposed for some LTI systems.

Now we will analyze the **time response** of systems to understand how much measuring systems distort the information contained in the measurand. For first-order LTI systems the **time constant** is a quality indicator of the dynamic response of the system; it is a way to characterize its dynamic response (http://www.sensorland.com). Equation (2.10a) represents a first-order LTI system. Recall that a linear time invariant (LTI) system is a system that can be described by a linear differential equation with constant coefficients,

$$\frac{d\,x(t)}{dt} + k\,x(t) = U(t) \tag{2.10a}$$

where $U(t)$ is the forcing function, or system input, and $x(t)$ is the system response, or system output.

As an example of a mechanical first-order LTI system we could consider the force $U(t)$ required to move a block of mass m at a velocity $x(t)$ sliding on an oil layer (Fig. 2.15a). The sliding viscous friction force is $f = b\,x(t)$, where b is the viscous friction coefficient. Then the differential equation is:

$$m\,\frac{dx(t)}{dt} + b\,x(t) = U(t) \tag{2.10b}$$

If $U(t)$ is the step function, the solution to Eq. (2.10a) is (Millman & Taub, 1965):

$$x(t) = C_0 + C_1\,e^{-t/\tau} \tag{2.11}$$

where τ is known as the time constant of the system and has units of time.

If $C_0 = 0$ the result is a decaying exponential, and when $C_0 = -C_1$ the solution can be written as

$$x(t) = -C_1(1 - e^{-t/\tau}) \tag{2.12}$$

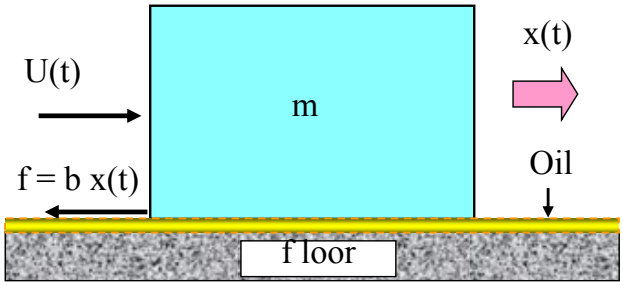

Fig. 2.15a: A block of mass *m* pushed with force *U*(*t*) and sliding at a velocity *x*(*t*) on an oil layer. It experiences a viscous friction force proportional to the velocity. The friction coefficient is b.

Figure 2.15b shows the functions $y_1 = x(t)/C_1$ for $C_0 = 0$, and $y_2 = - x(t)/C_1$ for $C_0 = - C_1$, both as functions of $z = t/\tau$.

For a system excitation with a step-like shape, the system response will rise or fall exponentially to approximately 63 % of its final (asymptotic) value when $t = \tau$. Table 2.3 presents some points of the y_1 and y_2 curves.

The **time constant** of a system was found to be a useful tool to evaluate the system frequency performance, because, as will be shown below it can be related to the system bandwidth. Using a step-like shape input signal, the **time constant** is easy to measure for some first order linear networks and could be used to estimate the bandwidth.

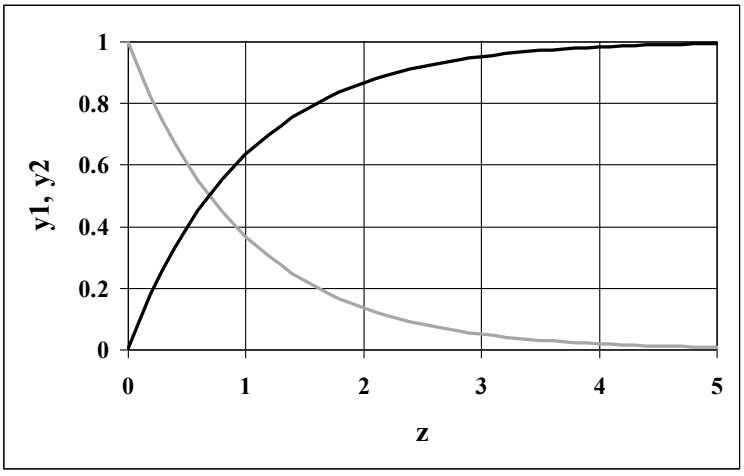

Fig. 2.15b: Solutions to Eq. (2.10a) when *U*(*t*) is the step function. $y_1 = x(t)/C_1$ and $y_2 = - x(t)/C_1$ as functions of $z = t/\tau$.

Table 2.3: Some points of the y_1 and y_2 curves (Fig. 2.15b)

z	y_1	y_2
0.5	0.607	0.393
1	0.368	0.632
2	0.135	0.865
3	0.050	0.950
4	0.018	0.982
5	0.007	0.993

2.4.6 Rise Time and Fall Time

Before showing how to relate frequency and time indicators of dynamic response quality, another tool will be introduced to evaluate the time response of a system: the **rise time**, which for special cases can be related to the **time constant**. When an input step function is applied to a system, the **rise time** (*rt*) refers to the time required for the output to change from low to high specified values; typically, these values are 10% and 90% of the step height (*H*). Conversely, the **fall time** is the time required for the output to change from 90 to 10 %.

$$rt = t\,(h_2) - t\,(h_1) = t\,(0.9H) - t\,(0.1H)$$

It is simple to establish a relation between the rise time and the time constant for any first-order LTI system. It was found (Millman & Halkias, 1967) that for this case the $h_1 = 0.1$ H is reached after $t = 0.105\,\tau$ and the $h_2 = 0.9$ H after $t = 2.302\,\tau$; thus,

$$rt = (2.302 - 0.105)\,\tau \approx 2.2\,\tau \tag{2.13}$$

2.4.7 Time Constant and Bandwidth Relation

This relation is well known in electronic circuits and can be found in the literature (Millman & Halkias, 1967; http://en.wikipedia.org/wiki/Time_constant), and then it will not be developed here. It can be demonstrated (Millman & Halkias, 1967) that the upper frequency of a first-order LTI system bandwidth is related to τ by

$$f_2 = \frac{1}{2\pi\tau} \tag{2.14}$$

A very interesting conclusion can be drawn from this equation: by applying a step input to a first-order LTI system, its time constant can be estimated and the upper frequency of the **bandwidth** can be determined by Eq. (2.14). Thus, for a first-order

LTI system the time and frequency response can be linked. It means that measuring the time response, some information on the frequency response can be obtained.

Figure 2.16 illustrates the time response of a temperature sensor. It was immersed in a bucket with soil at T = 20 °C until the thermal equilibrium was reached; then the soil temperature was suddenly changed (as a step-like shape input signal) to 46°C and the temperature change registered until the final temperature was reached. Because the time constant is the elapsed time since the change in temperature was initiated, until the 63.2 % of the total change was reached (Tab. 2.3), the temperature value to calculate the time constant is $T(\tau) = 20\ °C + 0.632 \times (46 - 20)\ °C = 36.432\ °C$, which correspond to $t_1 = 14$ s. Since the temperature change was initiated at $t_0 = 4$ s, the time constant is $\tau = t_1 - t_0 = 10$ s.

Once the time constant is estimated, the upper frequency of the bandwidth can be calculated from Eq. (2.14), the result being $f_2 = 0.016$ Hz. When using this sensor in a measuring application, Table 2.2 can be used to know the attenuation and phase change for a given frequency of the input temperature.

For example, applying the rule of thumb which says that to have low attenuation, the upper frequency of the system bandwidth should be one decade up the maximum frequency of the signal input, a phenomenon with a maximum frequency of 0.0016 Hz (period $T = 625$ s), measured with this temperature sensor, will result attenuated less than 0.5 %. In other words, sinusoidal temperature changes of about 10 minutes of period would be attenuated less than 0.5 %, but sinusoidal changes of about 1 minute would be attenuated about 30% (because 1 minute is approximately the period of the upper frequency $f_2 = 0.016$ Hz).

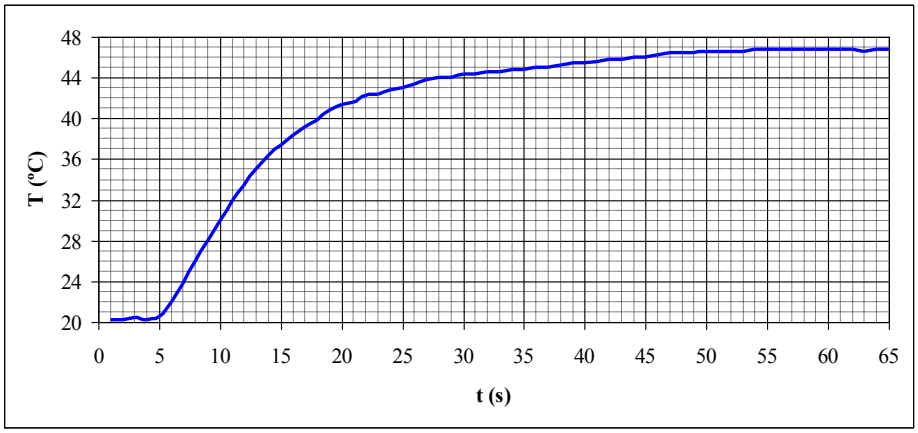

Fig. 2.16: Time response of a temperature sensor when a temperature step of 26 °C is applied to it. The delay from the beginning of the step until the temperature of the sensor reaches 63.2 % of the step amplitude is the time constant of the sensor.

2.4.8 Rise Time and Bandwidth Relation

Multiplying Eq. (2.13) by Eq. (2.14) we get,

$$rt\, f_2 = \frac{2.2\,\tau}{2\pi\tau} = 0.35 \tag{2.15}$$

The resulting product is a dimensionless constant that is often used to evaluate the bandwidth of a first-order LTI system by measuring the rise time.

2.4.9 Measuring the Rise Time of a Phenomenon by Means of an Instrument

The rise time of a phenomenon (d) that is to be measured by an instrument will be overestimated due to the own instrument rise time (i) according to Eq. (2.16), where m is the measured rise time (Walter, 2004).

$$m = \sqrt{(i)^2 + (d)^2} \tag{2.16}$$

A practical rule is to use an instrument whose rise time is 1/3 to 1/5 the rise time of the measured signal. In these cases errors introduced by the instrument are 5.5 and 2 % respectively. Therefore, if a steep slope of the phenomenon should be measured, it should be very important to verify that the rise time of the instrument meets Eq. (2.16).

2.4.10 Summary

Manufacturers inform the time and frequency characteristics of systems, instruments and sensors in diverse ways. When the information is reported in the frequency domain, the **Frequency response or Bode plots** permit the input-output relationships for any kind of systems to be described. **Bandwidth, upper and lower frequency and gain** allow the manufacturers to inform the dynamic properties of LTI systems.

When the information has to be reported in the time domain, the **rise and fall time** may be used for describing any type of systems. When the system is a first-order LTI one it is possible to compare the **rise and fall time** with the increasing and decreasing **time constants** of the system. Also, for these cases, **rise time** and **time constant** may be related to **bandwidth.**

Users should be aware of all these definitions to be able to understand the potential and limitations of the instruments they are using. They should confront the expected dynamic behavior of the measurand to the dynamic response of the systems to assure that the instrument does not significantly modify the parameter under study.

2.4.11 Examples to Help Fix Previous Concepts

Some examples extracted from real cases and somewhat modified will be presented to clarify the previously developed concepts These examples may be considered somewhat unrealistic but they were chosen in order to stress how in some cases the dynamic responses of the instruments could distort the measurand.

Example 1 - Figures 2.12a and 2.12b show the gain and phase (**frequency response**) as a function of **f** for a sea wavemeter whose sensor is an accelerometer; these figures are referred to as the **Bode plot** of the instrument response.

It can be draw from Figure 2.12a that this particular device has a maximum transference at $f \approx 0.8$ Hz, being G (0.8 Hz) ≈ 1.8 and φ (0.8 Hz) ≈ 9°. This means that a measurand (sea wave) of such frequency (0.8 Hz) will be measured by the instrument 80 % higher and with a phase lag of 9°. Wave energy is directly related to the square of the wave amplitude. Thus, because $A^2 = 1.8^2 = 3.24$, energy of this frequency component would be estimated with a considerable error.

Wave amplitudes whose frequencies are between 0.1 and 0.4 Hz are not modified, but phases are; for example φ (0.1 Hz) ≈ 30°. This behavior implies that if the study of the ocean is devoted to obtain wave energy between 0.1 and 0.4 Hz, because energy is related to the amplitude, results will be adequate. But if we are interested in the shape of the wave it will be distorted compared to the input shape due to the phase shifts.

For example, using the instrument between 0.1 and 0.4 Hz, where amplitudes are not modified, an artificial wave was synthesized by adding four sine waves of frequencies 0.1, 0.2, 0.3 and 0.4 Hz of the same amplitude and zero phase shift (Fig. 2.17, Input); the same signals were shifted in phase 30, 15, 10 and 9° respectively, which is approximately the phase lag indicated in the Phase Bode plot, and summed up (Fig. 2.17, Output). Clearly the shape has been drastically modified just due to the phase shifts. Perhaps in sea wave studies it could not be of importance to know the shape of the waves, but for other kinds of instruments it could be of great concern.

Suppose that two instruments from different manufacturers are used to analyze the traveling time of a wave between two points, one without phase shifts and other with a phase distortion similar to that of Figures 2.12a and 2.12b. Let us assume that such time is calculated by means of a correlation between the waves measured at both points. Since correlations are based on shape information, the calculated time could have very important errors because the instruments delay frequency input in different ways.

Finally, below 0.1 Hz the signal of the wavemeter results attenuated and the energy would be underestimated.

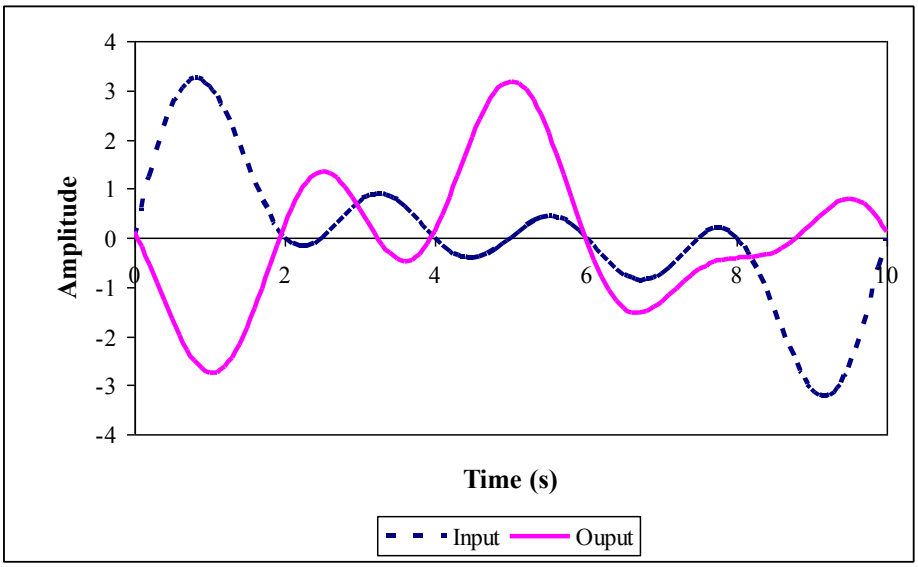

Fig. 2.17: An artificial wave was synthesized by adding four sine waves of frequencies 0.1, 0.2, 0.3 and 0.4 Hz and applied to the wavemeter dynamic transfer function of Figures 2.12a and 2.12b. Output shows how the shape of the wave is distorted due to the phase shift introduced by the meter.

Example 2 - What is the price paid when frequency components of the measurand equals the **bandwidth** of the instrument?

It has already been stated that some manufacturers specify the frequency response of instruments by means of their bandwidths. It was shown that measurand frequencies close to the upper and lower frequency of the transference will appear attenuated and shifted at the output.

Usually when users are buying an instrument they only verify that the bandwidth of the phenomenon being measured matches the bandwidth of the instrument. Let us see what happens with a signal whose bandwidth matches the instrument bandwidth. Suppose the instrument has f_1 =10 Hz, f_2 = 1000 Hz and the input signal (measurand signal) is the sum of three sine waves whose frequencies are: 10 Hz, 100 Hz and 1000 Hz. Figure 2.18 shows the input and output signals as function of time. Again, it is shown how the instrument could distort the shape of the measurand even if the bandwidth of the instrument matches the bandwidth of the phenomenon being measured. Perhaps for many applications this distortion is not a trouble, but for others it could be inconvenient. Users should be aware on how instrument characteristics are reflected in the recorded data.

Example 3 - This example illustrate how to evaluate whether an instrument of unknown or suspected frequency response can be used for a specific study. Suppose that a laboratory has a submergible pressure sensor with unknown specifications and

it is desired to know whether it could be used to measure waves with periods between 1 and 20 s.

The first step is to estimate the sensor's frequency response. It can be easily done if the sensor could be treated as a first-order LTI system.

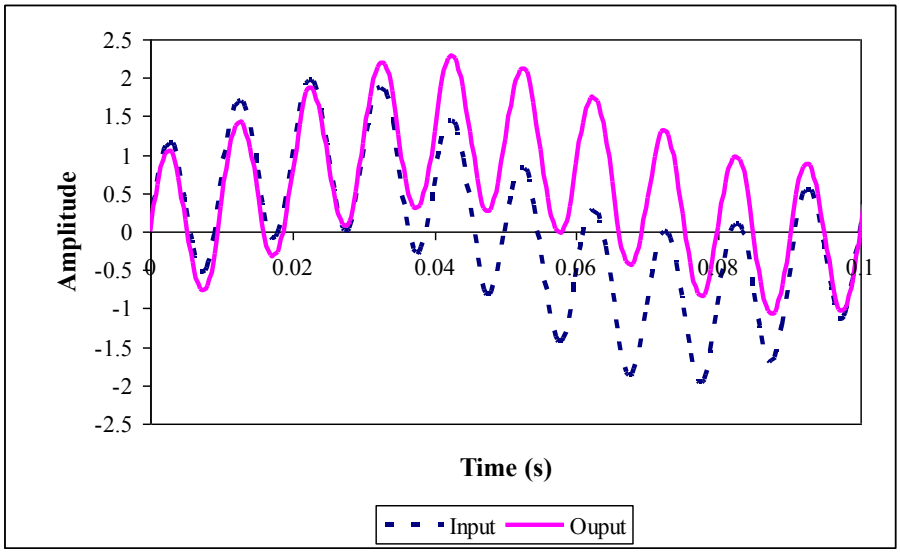

Fig. 2.18: An instrument has lower frequency f_1 =10 Hz and upper frequency f_2 = 1000 Hz, and the measurand signal is the sum of three sine waves whose frequencies are: 10 Hz, 100 Hz and 1000 Hz. Output signal is different than the input due to attenuation and phase shift introduced by the instrument.

In general, pressure sensors used in environmental sciences are able to measure the still water level; in other words, the lower frequency of the bandwidth is f_1 = 0. Then, if correctly installed, the sensor will measure long waves without attenuation.

In contrast, pressure sensors do have an upper -3 dB gain frequency and the higher frequency content of the measurand may result attenuated. Although in this case f_2 is unknown, some test can be done to estimate it. We have pointed out that the time constant gives information about the bandwidth of the instrument. Therefore a test to measure the time constant should be performed. For this purpose a pressure step input should be applied to the sensor.

In order to apply a step of pressure to the sensor it can be placed inside a pressurized balloon. Then the electrical output of the pressure sensor has to be connected to a data logger. After starting data recording the balloon has to be popped. Data have to be continuously recorded until the new pressure output becomes constant. For many

practical purposes this sudden change in pressure behaves like a negative step that allows the time constant to be measured. Proceeding as was done with Figure 2.16 the time constant of this hypothetical test was found to be $\tau = 10$ ms, then

$$f_2 = \frac{1}{2\pi\tau} = \frac{1000}{20\pi} Hz \approx 16\ Hz$$

Following the rule of thumb that states that for the input signal to pass the system with only 0.5% attenuation, the maximum frequency of the signal input should be one decade below the upper frequency of the bandwidth, this sensor could be used to measure from the still water level up to 1.6 Hz ($T = 0.625$ s). Therefore, it has been confirmed that this sensor can be used to measure waves with periods between 1 and 20 s if 0.5% of attenuation in the higher frequencies is accepted.

A more realistic and complete experiment to obtain time and frequency characteristics of pressure sensors are presented in Section (11.3). The ideas behind this experiment can be extended to other kind of sensors.

Example 4 - This example will show how far an instrument can follow a rapid change in the measurand without appreciably distorting the data. In this case the measurand is the wind speed and the instrument is an anemometer. It is desired to know how the time response of the instrument affects the measurement of fast changes in wind speed.

In general, mechanical anemometers are not first-order LTI systems because they do not respond in the same way to increasing and decreasing wind speeds. But in order to approximately know the response of an anemometer an increasing step of wind was applied to it and its output recorded. From the time data series the time required to reach 63.2 % of the final value was estimated and hence the time constant was established as $\tau = 2.5$ s.

Question: What is the minimum wind rise time (steep slope of the wind speed) that this instrument could measure if an error of 10 % is accepted?

Suppose that this anemometer behaves as a first-order LTI system, then following the naming of Eq. (2.16), the instrument rise time is $i \approx 2.2\,\tau = 2.2 \times 2.5$ s = 5.5 s. Since the maximum value admitted for the error of the measured rise time (m) is only 10 % larger than the value of the unknown rise time to be measured (d), we have $m \leq 1.1\,d$.

The minimum d for this condition may be calculated from Eq. (2.16),

$$(m)^2 = (i)^2 + (d)^2 \text{ and because } m \leq 1.1\,d;\ (1.21 - 1)d^2 = i^2;$$

$$\text{then } d \leq \sqrt{\frac{i^2}{0.21}} = 12\,s \tag{2.17}$$

Therefore, if the error in the slope of the wind speed must be smaller than or equal to 10 %, the minimum wind rise time that can be measured with this anemometer is 12 s. This means that for a change in wind speed that jumps from 10 to 90 % in less than 12 s, the instrument will overestimate the wind rise time in more than 10%.

For example, if d = 8 s, from Eq. (2.16) m = 9.7 s. It means that a wind rise time of 8 s will be measured as 9.7 s. The percentage of error can then be calculated as

$$e\,(\%) = \frac{(9.7-8)}{8} \times 100 \approx 21\%$$

2.5 Filters

Some general concepts on filters and filtering will be introduced in this chapter because as it was mentioned in the topic dedicated to time and frequency characteristics of sensors and systems, all stages of instruments are to some extent filters.

Electronic filters are networks specifically designed to process signals in a frequency-dependent manner (http://www.analog.com). Figure 2.19 shows the gain as a function of frequency for four ideal types of filters. The frequency response of a filter defines the filter properties. A **high pass filter** (HPF) is a filter that prevents passing low frequency inputs to the output. A **low pass filter** (LPF) allows passing low frequencies, but beyond some frequency the output decreases to zero.

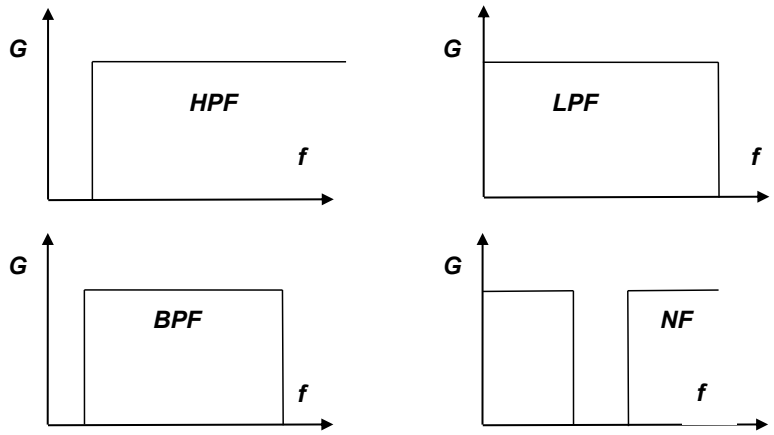

Fig. 2.19: HPF (high pass filter), LPF (low pass filter), BPF (band pass filter), NF (notch filter).

Connecting in series a low pass filter and a high pass filter a **band pass filter** (BPF) is obtained. BPF filter out low and high frequencies, keeping the central frequencies. A filter with the opposite characteristics than the BPF is known as a **notch filter** (NF). This last kind of filter is used to reject a particular range of frequencies. In the case of noise due to the AC power supply (for example f = 50 Hz) the notch is centered at this frequency.

Sensor manufacturing has technological limitations that make them inherently frequency selective and therefore acting similarly to filters. All devices produced by humans have limited time responses, or in other words, they have always a maximum frequency from which the sensor output begins to decrease until it becomes zero. This gives to sensors and then instruments an intrinsic low pass filter characteristic.

Because sensors are the first link of several elements in series that comprise the overall measuring instrument, then the sensor characteristics strongly influence the characteristic of the entire instrument. In general, all the information lost or distorted by the sensor cannot be recovered in the following steps. Digital signal processing helps to rescue signal buried in noise, but cannot substitute the quality of a sensor or save their weakness. Signal processing efforts should be made starting from a good sensor, thus the same effort will produce better results. The knowledge of the frequency response of the sensor allows knowledge of the "window" through which the sensor perceives the phenomenon being measured. Therefore, it is very important to select the adequate sensor for the desired measurand. The sensor should be selected to have a frequency response that allows passing the signals of interest without attenuation or distortion. A linear sensor that does not distort the signal must have constant gain and a linear phase change as a function of frequency, in all the measuring range of interest.

If the phenomenon to be studied is completely unknown, the frequency characteristics of the measurand could not be specified and this lack of information could lead to an inappropriate sensor or instrument. In this case, users should be aware that due to filtering the acquired information may be a limited version of the desired signal.

2.5.1 Noise Reduction by Filtering

As explained above, filtering is the property of a device to select some frequencies and reject others. It was mentioned that it is undesirable that a sensor filters out the input signal, but it would be desirable that some kind of filtering were used to reduce noise. Noise filtering is the process of removing undesirable noisy frequencies from signals. Filtering is a selective process that permits only useful information to be stored in, or transmitted by, the instrument, thus reducing the need for memory space or communication bandwidth.

The procedure of reducing the noise frequency components of a signal requires knowledge of the frequency characteristics of the wanted signals and the undesirable noise. By selectively rejecting some frequency bands, it is possible to extract the wanted signal from the noise, improving the signal to noise ratio. When the frequency content of the signal and noise are very different filtering is usually a simple solution. On the other hand, if both frequency of noise and input signal match, filtering can be extremely difficult and part of the input signal may be lost.

There are physical and mathematical filters; the first are analog circuits and are in general used in the first stages of an instrument or system. The second are applied after the signal has been converted from analog to digital. Most instruments incorporate microprocessors and **digital signal processing** (DSP) for a very quick mathematical manipulation of readings, thus making it possible that digital filtering of the signal be carried out on board the instrument.

Nowadays, some instruments allow users to decide how much filtering to perform before storing data. Some modern instruments have the capability of allowing users to select different input signal bandwidth, that is, permit tailoring the filter to their needs. In these cases, it is useful to identify the maximum frequency signal needed and keep the measuring bandwidth as low as possible. In this way high frequency noise is impeded from entering the system.

On the other hand, filtering modifies the data permanently, preventing a later signal processing on the original data. Designers have always a compromise on how much on board filtering will be applied, and users should make sure that filtering is not clipping the information they want.

When signals are low frequency and have random noise, averaging a large number of readings could produce good estimate of the true value. Present instruments use this technique to improve the data quality. Also, they usually calculate the standard deviation, which gives information on the amount of noise that accompanies the signal. The larger the standard deviation, the noisier is the signal. Some instruments verify that the standard deviation is below some specified level. If not, the data is discarded.

Several instruments record the signal to noise ratio and present this data to the users. This information can be employed to validate the data or not. In general, the manufacturers know and provide to the users the level of the minimum signal to noise ratio required to consider a data input as reliable.

A simple example of digital filtering follows: assume that air temperature data is needed for agricultural purposes and a sample of the soil temperature every ten minutes is enough. Then, if a data logger with a maximum sampling rate of one sample per second is available, a good solution to decrease random noise could be taking one sample every second and averaging 600 samples. This would give a good estimate of the soil temperature measured in the last 10 minutes. This averaged result will be then recorded by the data logger.

If the standard deviation (STD) is also recorded, and it was found in normal functioning conditions (for example during laboratory tests or tests under controlled conditions) that STD = 0.5 °C. This value could be used to validate field data, for example, disregarding data whose STD > 1 °C.

Let us see a different example in order to show how the dynamic of the measurand and the instrument performance condition the filtering process. If it is desired to study how a fan refrigerates a small metal piece, perhaps at least one temperature measure every one second will be required. Then, if the same data logger were used

(with maximum sampling rate of one sample per second), the averaging procedure to reduce noise described above could not be used. Therefore, the same data logger would produce more accurate readings for agricultural purpose than for metallurgical applications.

From these examples we can draw a general conclusion: it is possible to reduce errors by averaging noise **if the dynamics of the phenomenon is sufficiently slow with respect to the data acquisition rate**.

2.5.2 Filter Delay

In the previous example random noise was reduced by averaging 600 samples over 10 minutes; at the end of that period one value representing all these samples is recorded. Thus, averaging introduced a lag of 10 minutes to the data being measured. In the agricultural application, this delay in knowing the data is irrelevant, but could not be the case in other applications.

Assume a tide gauge placed at the entrance of a harbor which is used to judge the request of ships to enter or leave the port. As it happens in shallow water ports small changes in the tide may be important in allowing or rejecting the ship request. Then, to minimize the influence of waves on the sea surface measurement it could be desired to filter the data by averaging several samples of the sea surface over a period of time. The longer the period the lesser the wave influence, but the longer the delay in having the information.

Let us assume that the tide gauge averages during 20 minutes, then, the averaged data obtained is not the tide at the moment the data was calculated but the average over the previous 20 minutes. But, in the last 20 minutes the tide changed, and depending on the tidal cycle and weather conditions, it could change an amount not acceptable for the application the data is intended for.

Summarizing, filtering that allows knowing better the mean values of a phenomenon by averaging random noise or undesirable high frequencies, may introduce an unacceptable lag on the needed information.

2.5.3 Spatial Filtering

An interesting signal processing method that can be applied to some measurement cases is the spatial filtering of the signal. The concept of spatial filtering is the same discussed in Section (2.3.3). In this case spatial filtering is a wanted fact in opposition to the unwanted filtering shown before.

In order to explain this idea, the previous case of a tide gauge in a harbor is taken again. Let us assume that instead of having one tide gauge in the entrance of the port; a number of level meters are installed, distributed in a certain way (the allocation

has to be studied according to the port shape, bathymetry, etc). All level meters are connected to a computer that can process the information collected by them.

Suppose that these level meters allow a Low Pass Filter of one minute to be applied to the level signal (it could be done for example by sampling ten times by second and averaging 600 samples), thus these level meters will introduce a delay of only one minute, but the output level measured by each instrument will vary ostensibly due to long waves.

If the spatial distribution of the tide gauges is such that, at the same instant, different gauges measure the sea surface level at different phases of the waves, then, at the same time, some meters will report the wave crest, other the wave trough, while others will report the mean value of the sea surface. Then the average of the tide gauge outputs will tend to compensate the surface changes introduced by waves. This average of several levels measured in a certain area can be called a spatial filtering of the tide; it has a great advantage with respect to the time average presented previously for the harbor application. Because the average of several measures performed by the tide gauges is calculated in a short time, then in our example the tide is known with only a one minute delay. This is an interesting fact in such cases where a lag is not admitted and the phenomenon has a spatial distribution that allows the allocation of several sensors which measure the same measurand with different phases (Cavalieri & Curiotto, 1979).

2.6 Summary

The front ends of instruments are usually known as sensors and they are devices that take information from a physical or chemical phenomenon and create or modify an electrical signal. It can be found that they are named in diverse ways such as transducer, sensor, active transducer, passive transducer and detector. Also, devices that are actuated by a form of power and supply another one are called transducer, active transducer and actuator. Most of the time, all these devices will be mentioned in the future chapters as sensor or transducer.

Sensors and transducers are characterized by their transfer function (or transference) which is a curve or equation representing the relationship between the input and the output. A transfer function may be obtained with constant inputs (static transference) or varying inputs (dynamic transference).

Several parameters are used by manufacturers for specifying the static transfer such as: range, dynamic range, hysteresis, calibration curve, linearity, offset and gain errors, drift, etc.

It is a little harder to specify the dynamic transference of a sensor or instrument and for representing them the frequency response, expressed as a diagram of amplitude relations and phase relations, also known as Bode plot is used. This is a very detailed way of expressing a transference that can be used with any kind of

device. The drawback in obtaining it is that devices have to be excited with each frequency in the complete frequency range of interest and their outputs related in amplitude and phase to the respective inputs.

A simplified way to characterize the dynamic transference of first order linear time invariant (LTI) systems is to specify their bandwidth. In these cases the frequency response in amplitude and phase are linked by known equations, thus, by knowing the upper and lower frequencies the transference can be obtained. As a rule of thumb, when instruments can be treated as LTI systems and the upper frequency of the system is one decade up the maximum frequency of the signal input and the lower frequency is one decade down the minimum frequency of the input, the amplitude attenuation at the output is less than 0.5 % and the phase shift less than 0.6 °.

In order to characterize the dynamic behavior of sensors and instruments in the time domain, some parameters like time constant, rise time and fall time have been defined. For LTI systems it is possible to easily relate the time constant and bandwidth, thus, by means of a test in the time domain which provides the time constant, some frequency characteristics can be inferred.

Sensors and instruments are natural filters of the signal we want to measure because they discriminate how the different frequencies comprising the input signal are allowed to pass to the output. The instrument designer may also build some filters to let some frequencies pass or be blocked with the purpose of reducing noise. They are known as high pass, low pass, band pass and notch filters.

References

Carr J. J., & Brown, J. M. (1998). *Introduction to Biomedical Equipment Technology*. Upper Saddle River, NJ: Prentice Hall.

Cavalieri L., & Curiotto, S. (1979). A fast-response shallow-water tide gauge. *Il Nuovo Cimento*, serie *1*, vol *2-C*, 273-287. Datawell (1980). http://rpsmetocean.com/products_services/pdfs/oceanographic/Waverider%20f.pdf

Millman, J., & Taub, H. (1965). *Pulse, Digital and Switching Waveforms*. New York: McGraw-Hill Book Company Inc.

Millman, J., & Halkias, C. C. (1967). *Electronic Devices and Circuits*. Tokyo (Japan): McGraw-Hill Book Company Inc.

Walter, P. L. (2004). *Shock and Blast Measurement - Rise Time Capability of Measurement Systems*. Engineering Faculty, Texas Christian University, Fort Worth TX (USA) – PCB Piezotronics, Inc., Depew, NY (USA).

http://www.merriam-webster.com/dictionary

http://digital.ni.com/public.nsf/allkb/084702CE98679BB886256CA3006752D7

http://www.sensorland.com/HowPage026.html

http://en.wikipedia.org/wiki/Time_constant

http://www.analog.com/library/analogdialogue/archives/43-09/EDCh%208%20filter.pdf